Available online at www.sciencedirect.com

ScienceDirect

www.compseconline.com/publications/prodclaw.htm

Computer Law & Security Review

Comment

Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling

Michael Veale [a,*], *Lilian Edwards* [b]

[a] *Department of Science, Technology, Engineering & Public Policy (STEaPP), University College London, London, United Kingdom*
[b] *Strathclyde Law School, University of Strathclyde, Glasgow, United Kingdom*

Keywords:
Automated decision-making
algorithmic decision-making
Right to an explanation
Right of access
General Data Protection Regulation

A B S T R A C T

The Article 29 Data Protection Working Party's recent draft guidance on automated decision-making and profiling seeks to clarify European data protection (DP) law's little-used right to prevent automated decision-making, as well as the provisions around profiling more broadly, in the run-up to the General Data Protection Regulation. In this paper, we analyse these new guidelines in the context of recent scholarly debates and technological concerns. They foray into the less-trodden areas of bias and non-discrimination, the significance of advertising, the nature of "solely" automated decisions, impacts upon groups and the inference of special categories of data—at times, appearing more to be making or extending rules than to be interpreting them. At the same time, they provide only partial clarity – and perhaps even some extra confusion – around both the much discussed "right to an explanation" and the apparent prohibition on significant automated decisions concerning children. The Working Party appears to feel less mandated to adjudicate in these conflicts between the recitals and the enacting articles than to explore altogether new avenues. Nevertheless, the directions they choose to explore are particularly important ones for the future governance of machine learning and artificial intelligence in Europe and beyond.

1. Background

In relation to a data subject, Article 22 of the General Data Protection Regulation (GDPR)[1] prohibits (with exceptions) any "decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". This right was ported to the GDPR from the Data Protection Directive (DPD) 1995 (arts 12(a) and 15),[2] and itself borrowed from early French data

* *Corresponding author.* Department of Science, Technology, Engineering & Public Policy (STEaPP), University College London, Boston House, 36–38 Fitzroy Square, London W1T 6EY, United Kingdom.
E-mail address: m.veale@ucl.ac.uk (M. Veale).

https://doi.org/10.1016/j.clsr.2017.12.002

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

[2] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to

protection (DP) law.[3] The intent of the 1995 provision was to respond to fears in the early days of digitisation that automated, and hence potentially inscrutable and unchallengeable, decisions might prejudice access to important facilities such as credit, housing or insurance. In practice, the provision was little known and largely unused. However since it was migrated to Article 22 of the GDPR with little substantive change, the right has become the subject of much academic attention[4] for its possible utility in curbing the power of complex, opaque and often invisible machine learning (ML) algorithms. Such systems commonly now make or, more often, support decisions of huge citizen and consumer importance in public and private sector domains such as criminal justice, welfare, taxation, search, marketing, entertainment and political opinion-making. Much concern has been raised in legal, policy and journalistic circles over whether such systems may create discriminatory, biased or unfair results.[5]

Art 22 is not a simple article to construe, being rife with exceptions and complications. The right is excluded if the decision is necessary for a contract, authorised by Member State law, or based on explicit consent. If the first or third exceptions apply, then minimum explicitly prescribed safeguards must be put in place. Furthermore if the decision is based on "special" categories of personal data (defined in art 9 of the GDPR and including sensitive data such as health, race and religion), then automated decision-making is only allowed on the basis of explicit consent or substantial public interest (usually where lives are at risk) and again, "safeguards" must be put in place. What these "safeguards" entail has become particularly controversial especially when considering if, as some have claimed,[6] a "right to an explanation" of how or why algorithmic system made a decision is implied or explicit in the GDPR.

Art 22 is not the only part of the GDPR to have been pressed into service to regulate the rise of algorithmic decision-making. Information and access rights in arts 13–15, again derived from a longstanding pedigree in the DPD but now interestingly tweaked, provide for the first time that data subjects must be informed of the very existence of automated decision-making, including profiling, in addition to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". What this "meaningful information" might entail, both in theory and practice, has again become a subject of considerable enquiry.[7]

Against the backdrop of this renewed global interest in art 22 and other parts of the GDPR as remedies with which to "enslave the algorithm",[8] the Article 29 Data Protection Working Party (A29WP)'s release of their draft guidance on "Automated individual decision-making and Profiling"[9] has been eagerly awaited. The document is wide ranging, and weightier (in a literal sense, by page count) than any other GDPR guidance yet published by the body. Included are the definitions of both automated decision-making and profiling; elaborations and analysis of the specific automated decision-making provisions in Article 22; as well as the more general provisions on profiling and automated decision-making elsewhere in the GDPR. In addition, specific issues on children and data protection impact assessments (DPIAs) are explored. Best practice recommendations and a reading list are annexed.

2. Implications for information and access rights

In an important paper, Wachter et al. claim the information and access rights in Section 2 of the GDPR only guarantee general and *ex ante* information around algorithmic systems rather than *ex post* information about how an automated decision related to a particular data subject's circumstances was generated.[10] This conclusion has been relatively controversial, particularly in relation to how much 'heavy lifting' is done by the new addition of the term "meaningful" in comparison to the DPD.[11]

Implicitly and without fanfare, the A29WP appears to align themselves with Wachter et al's view, by agreeing that the arts 13–15 right to "meaningful information about the logic involved" provides a "more general form of oversight", rather than "a right to an explanation of *a particular* decision" [italics original].[12] The information should consist of "simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision, without necessarily always

the processing of personal data and on the free movement of such data, OJ 1995 L 281/31.

[3] Lee A Bygrave, 'Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17 Computer Law & Security Report 17 at 17.

[4] Bryce Goodman and Seth Flaxman, 'European Union regulations on algorithmic decision-making and a "right to explanation" ' (ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 2016); Mireille Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era' in J Bus and others (eds.) *Digital Enlightenment Yearbook 2012* (IOS Press, 2012); Dimitra Kamarinou, Christopher Millard and Jatinder Singh, 'Machine Learning with Personal Data' (2016) Queen Mary School of Law Legal Studies Research Paper No. 247/2016; Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76; Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law and Technology Review 18.

[5] See eg Campolo and others, *AI Now 2017 Report* (AI Now Institute 2017); Solon Barocas and Andrew Selbst, 'Big Data's Disparate Impact' 104 California Law Review 671.

[6] Goodman and Flaxman *op. cit.*

[7] Edwards and Veale (n 4).

[8] Lilian Edwards and Michael Veale, 'Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions" ' (Brussels Privacy Symposium, Vrije Universiteit Brussel, 2017). Available on SSRN: https://ssrn.com/abstract=3052831.

[9] Article 29 Working Party (A29WP), 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP 251, 3 October 2017). <https://perma.cc/3X54-2DGC>.

[10] Wachter and others *op. cit.*

[11] See eg Andrew Selbst and Solon Barocas, 'Regulating Inscrutable Systems', draft on file with authors; cf Andrew Selbst and Julia Powles, 'Meaningful Information and the Right to Explanation' (2017) 17 International Data Privacy Law <https://ssrn.com/abstract=3039125> accessed 10 December 2017.

[12] A29WP (n 9), section 2 at 24.

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002

attempting a complex explanation of the algorithms used or disclosure of the full algorithm".[13] Interestingly, earlier on (p 15) the A29WP also explicitly notes that art 15, which is triggered by a data subject explicitly seeking information, implicitly after processing has commenced, does not provide the data subject with more information than should have been provided under arts 13 and 14—articles which require a data controller, respectively, to provide such information at the time of processing or to the date subject if the data was not obtained directly from him or her.[14] Put together, this approach seems designed to fatally damage the chances of generating a personalised *ex post* "right to an explanation" from art 15(h) without severe judicial disagreement with these guidelines[15]—and indeed the example given of "meaningful information" at p 14 restricts itself to regurgitating back (i) input information provided by the data subject, (ii) relevant information provided by others (e.g. credit history) and (iii) relevant public information used in the decision (e.g. public records of fraud). In other words, A29WP has suggested that no information about the "innards" of the decision-making process—anything of a decompositional nature[16]—need be given. Another gap would be any information about how the training set was established, chosen, cleaned or so on. At this stage in the evolution of practice relating to algorithmic transparency, ignoring the training set is arguably both an overly restrictive and prescriptive approach.

Interestingly, in a different part of the document, namely, the "Good practice recommendations", the A29 WP still suggests that while "a complex mathematical explanation about how algorithms or machine-learning work" will generally not be relevant, it "should also be provided if this is necessary to allow experts to further verify how the decision-making process works".[17] At what point and how this optional provision of information becomes "necessary" is, unfortunately, not further pursued by the Working Party.

3. Implications for Article 22 definitions

Firstly, the A29WP comes down strongly in favour of Article 22 being read as a general prohibition rather than a right to opt-out.[18] The ambiguous language of this oddly worded provision has long been a subject of confusion,[19] with countries such as the UK opting to interpret the DPD as requiring notice in writing to trigger this 'right'.[20] Given that the language of the core provision is in essence unchanged, this could be seen as unauthorised law-making. It is worth noting however that this trend is not confined to the A29WP alone, as both the similar provision to the GDPR's Article 22 in the Law Enforcement Directive,[21] the data protection regime applicable to the police, and the UK's transposition of it in its draft Data Protection Bill,[22] also changes the provision from a right to a prohibition.

Whether a right or a prohibition, Article 22 is restricted in two ways: to 1) "solely" automated decisions; which 2) produce "legal" or "similarly significant" effects. Both of these concepts contain substantial ambiguity where guidance is welcome and the A29 WP provide this at pp 9–11.

3.1. *"Solely"*

In art 22, the definition of "solely" is crucial to the practical extent of the rights afforded to data subjects.[23] Many automated systems produce significant outputs about individuals e.g. relating to criminal bail, welfare benefits or potential for employment, but few do so without what is often described as a "human in the loop"—in other words they act as decision support systems, rather than autonomously making decisions. Indeed it is quite hard to think of many automated systems where *significant* decisions are made "solely" by algorithms—behavioural targeting of adverts being one possible example (though see below regarding whether such a use is "significant"), while financial products, which already exist within a highly regulated domain, are also commonly wheeled out as illustration. Yet if any human involvement at all is allowed, through literal interpretation, to exclude a system from the ambit of Article 22, then its reach will be small indeed. Worse still, it would be easy to introduce a nominal human into the loop, "rubber stamping" automated decisions in order to knock out art 22 rights. In fact there is some evidence that even where systems are explicitly intended only to support a human decision maker, for reasons of trust in automated logic, lack of time, convenience or whatever, then the system tends to de facto operate as wholly automated.[24] There is a strong argument therefore that rights to control "solely" automated decision making must also apply to decisions made with *some* degree of human involvement, though the extent of that degree is hard to set.

The A29WP provides two interesting statements here. Firstly, they note that "if someone routinely applies automatically generated profiles to individuals without any influence on the result, this would still be a decision based solely on automated processing". This implies that when considering if "solely" applies to an automated system, DPAs should consider how often the system operator disagree with the system outputs and changes or otherwise augments them. Looking forward, this would have interesting consequences. If the machine is claimed to *outperform* humans and treated as such, any human's involvement in the process designed to avoid the application of Article 22 this should necessarily be expected

[13] *Ibid.*, 14.
[14] *Ibid.*, 15.
[15] See eg Selbst and Powles *op. cit.*
[16] Edwards and Veale (n 4).
[17] *Ibid.*, 29.
[18] A29WP (n 9), 9.
[19] See eg Bygrave *op. cit.*
[20] Data Protection Act 1998, s 12(1).
[21] Regulation (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ 2016 L 119/89, art 11(1).
[22] Data Protection HL Bill (2017–18) (66), cl 47.
[23] *Ibid.*, Bygrave *op. cit.*; Wachter and others *op. cit.*
[24] Linda J Skitka and others, 'Accountability and automation bias' (2000) 52 International Journal of Human-Computer Studies 4, 701.

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002

to be effectively nominal. This would mean the system should be regarded as "solely" automated, and where the significance criterion is also met, will require a human system to exist in parallel.

The A29WP also adds that "meaningful human input" is required rather than a "token gesture" for the system to be categorised as not "solely" automated. This second perspective focusses on ensuring the human has, in the words of the Working Party, the "authority and competence" to change the decision. This forms an interesting challenge. It has been noted that where humans are involved in decision-making, they are often in "moral crumple zones", socially and culturally responsible for the errors of complex systems even where upon careful analysis blame is much harder to assign.[25] As and if machine systems become better at given tasks, we can expect maintaining non-token "authority and competence" to be a significant social and organisational challenge, further reducing the scope of avoiding Article 22 obligations.

Given these quite tricky sociotechnical issues, how will "solely" be assessed? A Data Protection Impact Assessment (DPIA) seems the obvious choice, yet Annex 2 of the recent A29WP DPIA guidance providing "Criteria for an acceptable DPIA" omits any mention of Article 22 rights and obligations, potentially creating confusion on the ground.[26] How this expanded notion of "solely" could practically be assessed from the point of view of the data controller or the data subject is one of the significant grey areas this guidance leaves in its wake.

3.2. *"Legal" or "similarly [significant]" effects*

The second main restraint on user rights over automated decision-making and profiling is whether a decision has legal effects or if not or, alternately, is "significant". While legal effects are fairly clearly restricted to cases where legal status is altered or legal duties created (e.g. assessment of immigration status; authentication of a legal contract) "significant" effects are much vaguer.

The A29WP suggests that "significant" decisions include those the potential to "significantly influence the circumstances, behaviour or choices of the individuals concerned", as well as those that may lead to individuals' "exclusion or discrimination".[27] The use of "influence" not just "cause" interestingly suggests that systems that "nudge" individuals e.g. by changing the way that the choices they have available are presented may fall within these provisions, even if the final decision is left at least ostensibly to the individual—a discussion familiar to those following the literature on profiling and the law.[28]

Connectedly, systems generating differential pricing according to the profiled characteristics of a data subject ("price discrimination"[29]) would also be considered significant "if, for example, prohibitively high prices effectively bar someone from certain goods or services".[30] Interestingly, according to the A29WP, significant effects can be positive or negative. This does not explicitly depart from the 1995 Directive, but interestingly it does depart from early drafts which only restricted decisions "adversely" affecting individuals.[31]

A key issue on which views have differed since the 1995 Directive is automated targeting of adverts can ever be "significant".[32] On the one hand this is one of the most ubiquitous experiences of "solely" automated decision-making and the mismatch of targeted adverts with user expectations is a prime source of distrust of profiling in general.[33] On the other hand, adverts are not commands—it is arguable they are not even "decisions"—and can be easily ignored or blocked, even though they may shape individual experiences over time. The A29WP takes a middle line, suggesting that adverts targeted on simple demographics such as gender, age or city, do not have a "significant effect" on the recipient but that some adverts may, depending on:

- the intrusiveness of the profiling process;
- the expectations and wishes of the individuals concerned;
- the way the advert is delivered; or
- the particular vulnerabilities of the data subjects targeted.

These are mostly sensible, obvious factors to pick out—there is strong consensus, to pick two examples, that targeting anorexics with emetics, or profiling using recorded overheard conversations rather than disclosed text are unsavoury practices—but do these matters make the decision more *significant* or just more unpleasant? In the former case, perhaps yes—but what about the latter? There is a danger here that the A29WP guidance, well-meaningly, is drawing more from consumer protection principles than the underlying text.[34] On the other hand, it is odd to see no mention of the theory that targeted adverts are significant because, like price discrimination mentioned immediately below, they reduce the universe of

[25] Madeline Claire Elish, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction' (2016) We Robot 2016 Working Papers, doi:10.2139/ssrn.2757236.

[26] Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679' (WP 248 rev.01, 4 October 2017), 22.

[27] A29WP (n 9), 10,

[28] See generally Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (2015, Edward Elgar).

[29] See Frederik Zuiderveen Borgesius and Joost Poort, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 Journal of Consumer Policy 3, 347–366.

[30] The Working Party do not consider as some scholars have that a higher price could be an "invitation to enter an agreement", and thus have potential legal effect. See *ibid*., 362.

[31] Bygrave *op. cit*.

[32] Isak Mendoza and Lee A Bygrave, 'The Right not to be Subject to Automated Decisions based on Profiling' in Tatiana Synodinou and others (eds.) *EU Internet Law: Regulation and Enforcement* (Springer, 2017).

[33] For example, refer to the persistent speculation that FB "listen" to mobile users through their smartphone microphone and uses this to send ads related to conversations; FB denies this and it indeed seems unlikely it is necessary given the volume of other data and metadata they can draw on. See Zoe Kleinman, 'Facebook denies "listening" to conversations' *BBC News* (28 October 2017) <http://www.bbc.com/news/technology-41776215> accessed 14 November 2017.

[34] cf Natali Helberger, Frederik Zuiderveen Borgesius and Agustin Reyna, 'The Perfect Match? A Closer Look at The Relationship Between EU Consumer Law and Data Protection Law' (2017) 54 Common Market Law Review 1427.

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002

those deemed not suitable, not rich or not persuadable enough by certain offers and this create social sorting.

The issue of the significance of adverts dovetails into the overarching question of whether the effects must be "significant" to individuals, or if it suffices that they are significant to a group of which the data subject is a member.[35] For example, an advert targeted to those with "black-sounding" first names, suggesting that the aid of a criminal defence lawyer may be needed, does little to harm the reputation of the particular black, Harvard security professor, Latanya Sweeney, that was investigating the phenomenon when it occurred to her, but may arguably create a penumbra of racial bias and expectations of illegal behaviour around the entire group of black people, some of whom will be more vulnerable than our professor subject.[36] Alternately it could be argued that this is to confuse cause and effect; the group profile gave rise to the targeted advert which failed to significantly affect the professor. At root here is an irreconcilable tension between DP as a creature of the individual rights paradigm, and the inevitable conclusion that algorithmic decision-making leads to group harms. As above there is a choice to be made here between addressing social harms and letting equality law leak into DP law, or maintaining a more conservative separation.

The A29WP gives confusing signals in this respect. On one hand, they highlight that "[p]rocessing that might have little impact on individuals may in fact have a significant effect on certain groups of society, such as minority groups or vulnerable adults". Yet the example they then cite, of the vulnerable person in financial difficulties who is targeted with invites to online gambling, re-individualises the problem again to a person rather than a group (p 11): someone's identity as a gambling addict is defined primarily by their gambling behaviour, which is a quite different notion from someone whose group is defined by membership of a minority, and potentially protected, class – and thus the relation of protected characteristics to "significant" decisions remains unclear.

A paragraph or so later, however, we find the suggestion that the characteristics of the group may lead to detrimental algorithmic decisions about an individual data subject, in that significant effects "may also be triggered by the actions of individuals *other* than the one to which the automated decision relates" [italics added]. An example is given of where postcodes for down-at-heel areas might contribute to poor credit scoring for an otherwise creditworthy individual who lives there. There is no reason why such decisions should not fall within art 22—it is the decision that concerns the data subject that triggers it, even if the data used to *make* the decision comes partly or wholly from elsewhere. In fact such "peer related" factors are the norm rather than the exception in machine learning. And these are clearly the cases where transparency and associated opportunities for challenge under art 22 would be of the greatest use. In equality law proper, whether this use of group data to inform personalised decisions is "fair" is a conundrum which has been grappled with not always to produce optimum results, as with the famous decision of the CJEU[37] to proscribe discrimination in insurance results on the grounds of gender, which was expected to raise premiums for both sexes, but has been accused of further widening the gap between male and female quotes through the use of proxies.[38]

3.3. *Suitable safeguards*

3.3.1. *Updates on the "right to an explanation"*

The alleged "right to an explanation" safeguard in Article 22 has, as already noted, been the subject of both significant hope and contention, the latter particularly surrounding its legally ambiguous status[39] and its technical dimensions and practical use (or inherent limits) as a remedy.[40] Many of the issues relate to the inclusion of the right in the recitals of the GDPR, written in mandatory language, but its exclusion from the main article for political reasons, where a similar, shorter list of safeguards is provided.[41] The A29WP, clearly suffering themselves from this contradiction, does not address it head on, and therefore only compound the confusion it generates. In one page, they even manage to provide an Article 22-based safeguard list (omitting a right to an explanation) with the contradictory Recital 71-based list (including a right to an explanation) footnoted.[42] Furthermore, they emphasise that the Recital 71 safeguards apply "*in any case*" [italics in original].[43] They go on, in their "Key GDPR provisions that reference automated decision-making as defined in Article 22" section to choose the Recital 71 list and omit the Article 22 list,[44] whilst in the page preceding of "good practice" suggestions for suitable safeguards the right is conspicuous only by its absence, even in the presence of without explicit basis in recitals at all, such as "ethical review boards to assess the potential harms and benefits to society"![45]

The second set of issues, concerning the practical types of information that can be delivered by means of any Article 22-based explanation, are not addressed at all in the text. All references elaborating on algorithmic explanations relate to Section 2 information provisions rather than Article 22 safeguards, which as noted above, have a more general purview. Whether such a right to an explanation, were it to exist might include the kinds of "subject-centred" explanations (such as decision sensitivity to changes in input variables[46]), and how this right might play with the range of explanation facilities, both static and interactive, being pushed by computer

[35] Edwards and Veale (n 4).

[36] Latanya Sweeney, 'Discrimination in Online Ad Delivery' (2013) 56 Communications of the ACM 5, 44–54; *ibid.*

[37] Case C-236/09, Association belge des Consummateurs Test-Achats ASBL and Others v. Conseil des ministres [2011] ECR I-00773.

[38] Patrick Collinson, 'How an EU gender equality ruling widened inequality' *The Guardian* (London, 14 January 2017) <https://www.theguardian.com/money/blog/2017/jan/14/eu-gender-ruling-car-insurance-inequality-worse> accessed 14 November 2017. The use of proxy variables is well-studied in relation to fairness in algorithmically made and supported decisions. See eg Barocas and Selbst (n 5).

[39] Wachter and others *op. cit.*

[40] Edwards and Veale (n 4).

[41] Wachter and others *op. cit.*

[42] A29WP (n 9), 9.

[43] *Ibid.*, 16.

[44] *Ibid.*, 31.

[45] *Ibid.*, 30.

[46] For a broader typology, see Edwards and Veale (n 4).

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002

scientists in conferences and workshops such as Fairness Accountability and Transparency in Machine Learning (FATML) is yet to be seen.

3.3.2. *Children*

The A29WP was faced with an unenviable task in relation to automated decision-making and children: reconciling an absolute ban in the recitals with silence in the main text. Recital 71 states bluntly that the types of decisions in Article 22(1) "should not concern a child". Article 22, however, does not mention children. The Working Party's decided that it therefore "does not consider that this represents an absolute prohibition on this type of processing in relation to children" but that "wherever possible" controllers should not rely on Article 22(2) exemptions to justify it. This is in line with the interpretation of Mendoza and Bygrave,[47] who suggest that this will "likely increase the stringency" of the measures in Article 22, such as which decisions are construed as significant. The A29WP add to this to note that where this decision-making may need to be carried out (e.g. to "protect [children's] welfare") this must be alongside safeguards "appropriate for children". The trade-off made in this draft guidance, being value-charged as discussions around children's rights often are, has already led to heated debate in at least national legislature.[48]

3.3.3. *Discrimination-aware profiling, including machine learning*

The GDPR is quiet, although not silent, on bias and discrimination within algorithmic systems. The most direct allusion to it is found in a very long, winding sentence in Recital 71,[49] which notes the controller should "implement technical and organisational measures [. . .] in a manner [. . .] that prevents, inter alia, discriminatory effects on natural persons" on the basis of special categories[50] of data. This can be read in light of "discrimination aware" or "fairness aware: data mining and machine learning, a growing field of research and practice.[51] While "fairness" is an overarching principle of the GDPR, it is an extremely under-determined notion in data protection that has never been substantially attached to non-discrimination in processing outcomes.[52]

There are three main areas where the A29WP expresses views on discrimination and bias in machine learning systems used in profiling and automated decision-making.

First, they suggest measures to tackle discrimination as automated decision-making safeguards under Article 22(3–4). Following the allusion to discrimination-aware data mining in Recital 71, they recommend data controllers "design ways to address any prejudicial elements", "audit algorithms", and undertake "regular" and "cyclical" reviews to avoid discrimination on the basis of special category data. The enforceability of these will depend on whether the safeguards listed in Art 22 (only "the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision"), which are required "at least", are seen as minimally "suitable", or whether the balancing of safeguards against subjects' rights and freedoms will oblige further action.

Second, they suggest that data controllers relying on the legitimate interests grounds to justify profiling must particularly consider safeguards concerning "fairness, non-discrimination and accuracy".[53] Legitimate interest can never serve as grounds for the automated decisions considered in Article 22(1) so this can only apply to profiling where processing is not completely automated. It is unclear whether the "fairness" discussed in this section is linked in any way to the "fairness" principle in Article 5(1)(a) though this would seem a natural assumption.[54] A29WP nevertheless advises, albeit within an illustrative example, that within the information and access rights of arts 13–15, discussed earlier, controller should consider providing data subjects with information that their profiling methods "are regularly tested to ensure they remain fair, effective and unbiased".[55]

Thirdly, they refer to the obligations of data controllers who infer special categories of data from "ordinary" personal data through the use of profiling. This is a fraught area given increasing evidence that it is easy to derive from quotidian data such as social media posts and shopping bills, sensitive data about e.g. health and political opinions.[56] The A29WP argues this places a duty on such a data controller to notify data subjects not just that they have collected the data but that such sensitive inferences have been made (p 22). This has interesting logical consequences for the whole topic of discrimination aware ML. Seen at its most restrictive, it might imply that whenever a data controller can reasonably foresee they may "create" special categories of data i.e. in any case where non-sensitive personal data becomes a proxy for a special category of personal data in a model, such processing will have to take place

[47] Mendoza and Bygrave *op. cit.*

[48] HL Deb 13 November 2017, volume 785, cols 1865, 1870.

[49] Recital 71, GDPR notes that "the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures [. . .] that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect".

[50] What constitutes "special categories of personal data" in the GDPR is defined in Article 9(1). "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited".

[51] Dino Pedreshi, Salvatore Ruggieri and Franco Turini, 'Discrimination-aware data mining' (2008) Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08); more recently, see the proceedings of the four (as of 2017) Workshops on Fairness, Accountability and Transparency in Machine Learning (FAT/ML) at www.fatml.org.

[52] See Damian Clifford and Jef Ausloos, 'Data protection and the role of fairness' (3 August 2017) CiTiP Working Paper Series 29/2017.

[53] A29WP (n 9), 21.

[54] Clifford and Ausloos *op. cit.*

[55] A29WP (n 9), 14.

[56] See eg Svitlana Volkova and Yoram Bachrach, 'On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure.' (2015) 18 *Cyberpsychology, Behavior, and Social Networking* 12 (2015) 726–736.

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002

under the restricted grounds for processing of Article 9. This would, in most commercial cases, exclude the use of legitimate interests and require explicit consent to the creation of the sensitive data. In particular, it would place a particular imperative on data controllers to undertake prior analysis of the fairness characteristics[57] of their systems during the training period to observe if such proxy effects occur and if so, either to amend them or to seek new consents from the data subjects. This would be somewhat radical and provoke a refined emphasis on DP impact assessments which are also emphasised in the guidance.

4. Conclusion

This weighty guidance, despite some careful ambiguities in wording, leans at times nearer to unauthorised law-making than mere interpretation. This is particularly evident in the sections on what "solely", and "significant effects" mean in the context of art 22, as well as in the attitude taken to art 15 and its interaction with the "right to an explanation". On the other hand, little tangible help is given in relation to whether that elusive right *can* be derived from art 22 or recital 71, and (perhaps unsurprisingly) no help is given at all on what kind of elements might go into such explanations. A29WP might have referred the reader helpfully to the 2016 French Digital Republic Act[58] which gives quite detailed instructions on what information should be provided by way of explanation of algorithmic decisions in the administrative public sector, a model which has also been suggested as an amendment to the UK's Data Protection Bill in the course of its House of Lords process.[59] Another failure, albeit fairly understandable given the difficulties of the text, is to address head on the issue of whether and when profiling of children is allowed. Finally, there are interesting future hooks for regulators and (probably) national courts in relation to algorithmic bias and discrimination. The rules for inferred special categories of data are likely only to become ever more controversial as the deployment of political, racial and economic modelling of data subjects through casual online exchanges, clicks and "Likes" becomes more apparent.[60] Indeed, these rules, which were perhaps unexpected in their conviction and especially pertinent when considering the nature of data transformation in modelling, could end up very important indeed for future governance of profiling in Europe.

Acknowledgements

Michael Veale acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M507970/1]. Lilian Edwards was supported in part by the Arts and Humanities Research Council (AHRC) centre *CREATe* [grant number AH/K000179/1], and the EPSRC Digital Economy Hub Horizon at the University of Nottingham [grant number EP/G065802/1]. We thank two anonymous reviewers for their helpful comments on the manuscript, and Steve Saxby for his outstanding editorial responsiveness.

[57] Note that the special categories of data in the GDPR are not the same as most countries' "protected characteristics", for those that have them. Gender, for example, is not included as a special category in the GDPR, whereas political opinion is.

[58] Loi n 2016-1321 du 7 October 2016 pour une République numérique, art 3; for the information that must be provided to the citizen upon request, see Code des relations entre le public et l'administration, version consolidée au 1 September 2017, art L311-9. For English translation and analysis, see Edwards and Veale (n 8).

[59] HL Deb 13 November 2017, volume 785, cols 1863.

[60] See e.g. the 2016 case of Facebook allowing advertising targeting by "ethnic affinity": Julia Angwin and Terry Parris Jr, 'Facebook lets advertisers exclude users by race' *ProPublica* (28 October 2016) <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.

Please cite this article in press as: Michael Veale, Lilian Edwards, Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.12.002